THE ACCURACY OF MANTEL-HAENSZEL, SIBTEST, AND LOGISTIC REGRESSION METHODS IN DIFFERENTIAL ITEM FUNCTION DETECTION

Budiyono
Universitas Sebelas Maret
bud@uns.ac.id

Abstract

This simulation study was aimed at answering the questions of a) what is the order of accuracy among Mantel-Hanzel, Sibtest, and Logistic Regression methods in Differential Item Function (DIF) detection and b) what percentage of DIF that makes Mantel-Haenszel, Sibtest, and Logistic Regression methods not produce any error. There were 1000 participants for the reference group and 1000 participants for the focal group. There were three groups of test being studied namely 20 tests with the length of 25, 40 tests with the length of 50, and 60 tests with the length of 75. The DIF contained in the *i*-th test was 4i% for the first group, 2i% for the second group, and 4/3i% for the third group. The data is dicotomous generated by using Fortran language. Findings of the simulation suggested that a) ordered from most to least accurate, the order of the methods were SIBTEST, Mantel-Haenszel, and Logistic Regression, b) if the amount of DIF is 8% at most, the Mantel-Haenszel method did not have any error, SIBTEST did not have any error at any condition, and if the amount of DIF in a test were 24% at most in the first group, 8% at most at the second group, and 7% at most in the third group, the logistic regression method did not have any error.

Key words: *DIF, comparation of accuracy, Mantel-Haenszel, SIBTEST, logistic regression*

KETEPATAN METODE MANTEL-HAENSZAL, SIBTEST, DAN REGRESI LOGISTIK UNTUK MENDETEKSI *DIFFERENTIAL ITEM FUNCTION*

Abstrak

Penelitian dilakukan untuk mengetahui (a) urutan ketepatan pendeteksian *Differential Item Functioning* (*DIF)* antara metode Mantel-Haenszel, *the Simultanous Item Bias Test* (SIBTEST), dan regresi logistik, (b) persentase kandungan *DIF*, Mantel-Haenszel, SIBTEST, dan regresi logistik tidak melakukan kesalahan. Simulasi dipilih dengan desain sebagai berikut (a) peserta tes sebanyak 1000 orang untuk kelompok acuan dan 1000 orang untuk kelompok fokus, dan (b) tiga kelompok tes pilihan ganda dengan 5 pilihan jawaban. Ada 20 buah tes masing-masing terdiri dari 25 butir soal, 40 buah tes masing-masing terdiri dari 50 butir soal, dan 60 buah tes masing-masing terdiri dari 75 butir soal. Muatan *DIF* pada tes ke-i sebesar 4i% pada kelompok pertama, 2i% pada kelompok kedua, dan $\frac{4}{3}$i% pada kelompok ketiga. Hasil penelitian menunjukkan bahwa: (a) urutan ketepatan mendeteksi *DIF,* secara berturut-turut adalah metode SIBTEST, metode Mantel-Haenszel, dan metode regresi logistik, (b) pada kandungan *DIF* paling banyak 8%, metode Mantel-Haenszel tidak melakukan kesalahan, metode SIBTEST tidak melakukan kesalahan pada setiap persentase kandungan *DIF*, dan pada kandungan *DIF* paling banyak 24% (untuk panjang butir 25), 8% (untuk panjang butir 50), dan 7% (untuk panjang butir 75), metode regresi logistik tidak melakukan kesalahan.

Kata kunci: *DIF, perbandingan ketepatan, Mantel-Haenszel, SIBTEST, regresi logistik*

Pendahuluan

Suatu tes biasanya terdiri atas sejumlah butir soal. Tes yang baik harus terdiri atas butir-butir soal yang baik. Butir soal yang baik, antara lain, harus mempunyai tingkat kesulitan yang memadai dan mempunyai daya pembeda yang baik. Selain itu, untuk butir soal yang baik pada tes pilihan ganda, pengecoh yang disediakan harus dipilih oleh sebagian peserta tes yang kemampuannya rendah.

Sejak tahun 1960-an, kecuali persyaratan-persyaratan yang telah disebutkan, terdapat persyaratan tambahan agar butir soal dikatakan baik, yaitu butir soal harus adil (*fairness*). Pengujian untuk melihat apakah butir soal bertindak adil atau tidak disebut pengujian *Differential Item Functioning* (*DIF*). Secara konseptual, *DIF* dikatakan muncul pada sebuah butir soal jika peserta tes yang mempunyai kemampuan yang sama pada konstrak yang diukur oleh tes, tetapi berasal dari kelompok berbeda, mempunyai peluang berbeda dalam menjawab benar butir soal tersebut (Hulin, Drasgow & Parson, 1993: 152).

Penentuan apakah suatu butir soal terindikasi *DIF* atau tidak memerlukan indeks *DIF*, yaitu indeks yang menunjukkan seberapa kuat indikasi *DIF* ada pada butir itu. Jika tingkat indikasi *DIF* tersebut secara praktik dianggap signifikan, dapat dengan mengujinya memakai uji statistik tertentu atau hanya dengan melihat indeksnya saja, maka butir soal yang bersangkutan dikatakan terkena *DIF*, memuat *DIF*, atau terdeteksi sebagai butir *DIF*.

Terdapat dua jenis *DIF*, yaitu *DIF* uniform (konsisten) dan *DIF* tidak uniform (tidak konsisten). *DIF* uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan *DIF* tidak uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan (Penfield & Lam, 2000: 9).

Ada beberapa metode pendeteksian DIF yang banyak dipakai dalam praktik pengukuran dan pengujian dewasa ini. Metode-metode itu adalah Metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik.

Pada penggunaan metode Mantel-Haenszel, peserta tes pada setiap kelompok (kelompok fokus dan kelompok acuan) digolongkan menjadi M buah kategori berdasarkan pada level kemampuan peserta tes. Kemampuan peserta tes ini disebut variabel pemadanan, yaitu variabel yang dipakai sebagai dasar untuk pemadanan (Holland & Thayer, 1993: 39). Pada metode Mantel-Haenszel, kemampuan peserta tes diwakili oleh skor total peserta tes. Data yang digunakan dalam metode Mantel-Haenszel adalah data pada tabel kontingensi 2×2 sebanyak M buah atau data pada sebuah tabel kontingensi besar berukuran 2×2×M, dengan M adalah banyaknya penggolongan atas dasar level kemampuan peserta tes. Setiap tabel kontingensi 2×2 berbentuk seperti pada Tabel 1.

Tabel 1. Tabel Kontingensi 2 × 2 untuk Butir Soal Tertentu pada Level Kemampuan ke-m

	Banyaknya Peserta Tes yang Menjawab Benar	Banyaknya Peserta Tes yang Menjawab Salah	Banyaknya Peserta Tes Secara Keseluruhan
Kelompok fokus (f)	R_{fm}	W_{fm}	N_{fm}
Kelompok acuan (r)	R_{rm}	W_{rm}	N_{rm}
Kelompok total (t)	R_{tm}	W_{tm}	N_{tm}

Hipotesis nol *DIF* pada metode Mantel-Haenszel yang diuji secara statistik adalah:

$$H_0 : \frac{R_{rm}}{W_{rm}} = \frac{R_{fm}}{W_{fm}}, \text{ untuk } m = 1, 2, 3, \ldots, M \quad (1)$$

Hipotesis alternatif yang bersesuaian dengan hipotesis nol tersebut adalah:

$$H_a : \frac{R_{rm}}{W_{rm}} = \alpha \frac{R_{fm}}{W_{fm}} \quad (2)$$

untuk $m = 1, 2, 3, \ldots, M$ dan $\alpha \neq 1$

Parameter α pada (2) disebut *common odds ratio* pada M buah tabel kontingensi 2×2, sebab di bawah H_a, nilai α adalah *common odds* untuk setiap m, yaitu:

$$\alpha_m = \frac{\frac{R_{rm}}{W_{rm}}}{\frac{R_{fm}}{W_{fm}}} = \frac{R_{rm} W_{fm}}{R_{fm} W_{rm}} \quad (3)$$

Mantel dan Haenszel menyediakan estimasi untuk *common odds ratio* sebagai berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$\hat{\alpha}_{MH} = \frac{\sum_m \frac{R_{rm} W_{fm}}{N_{tm}}}{\sum_m \frac{R_{fm} W_{rm}}{N_{tm}}} \quad (4)$$

Jika $\hat{\alpha}_{MH} > 1$, maka butir yang diselidiki terkena *DIF* yang menguntungkan kelompok acuan. Jika $\hat{\alpha}_{MH} < 1$, maka butir yang diselidiki terkena *DIF* yang menguntungkan kelompok fokus.

Uji signifikansi hipotesis nol $H_0 : \alpha_m = 1$, untuk setiap m, menggunakan statistik tes khi-kuadrat sebagai berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$MH_{\chi^2} = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - 0{,}5\right]^2}{\sum_m Var(R_{rm})} \quad (5)$$

dengan

$$E(R_{rm}) = E(R_{rm} \mid \alpha = 1) = \frac{N_{rm} R_{tm}}{N_{tm}} \quad (6)$$

$$Var(R_{rm}) = Var(R_{rm} \mid \alpha = 1) = \frac{N_{rm} R_{tm} N_{fm} W_{tm}}{N_{tm}^2 (N_{tm} - 1)} \quad (7)$$

Statistik uji MH_{χ^2} pada persamaan (5) terdistribusi menurut distribusi chi-kuadrat dengan derajat kebebasan 1, jika H_0 benar. Kriteria pengambilan keputusannya adalah sebagai berikut. Jika $MH_{\chi^2_{obs}} > \chi^2_{\alpha;1}$, maka butir soal yang bersangkutan secara signifikan terdeteksi *DIF*.

SIBTEST (*the Simultaneous Item Bias Test*) merupakan sebuah alternatif metode statistik untuk mendeteksi *DIF* yang dikemukakan oleh Shealy dan Stout (1993: 198). Pada metode SIBTEST, keseluruhan butir soal dibagi menjadi dua subtes, yaitu *the studied subtest* atau *the suspect subtest* dan *the matching subtest*. *The studied subtest* berisi butir atau butir-butir soal yang diselidiki *DIF*nya, yang untuk selanjutnya disebut subtes yang diselidiki. *The matching subtest* berisi butir-butir soal sisanya, yang untuk selanjutnya disebut subtes pemadanan. Subtes pemadanan dipakai untuk mengelompokkan kedua kelompok (yaitu kelompok fokus dan kelompok acuan) ke dalam sub-sub kelompok yang komparabel pada kemampuan yang diukur.

N adalah banyaknya butir soal. Contoh, butir-butir 1, 2, …, n menyatakan subtes pemadanan, dan butir-butir n+1, n+2, …, N menyatakan subtes yang diselidiki. Untuk setiap peserta, $X = \sum_{i=0}^{n} U_i$ menyatakan skor total pada subtes pemadanan dan $Y = \sum_{i=n+1}^{N} U_i$ menyatakan skor total pada subtes yang diselidiki. Semua peserta dalam kelompok acuan dan kelompok fokus dikelompokkan ke dalam K subkelompok (maksimum K = n+1) berdasarkan skor mereka pada subtes pemadanan. Peserta pada kelompok acuan dan kelompok fokus yang mempunyai skor subtes pemadanan yang sama dibandingkan. Selisih rerata terbobot antara kelompok acuan dan kelompok fokus pada subtes yang diselidiki pada seluruh K subkelompok diberikan oleh (Hsin-Hung Li, Nandakumar & Stout, 1995: 5):

$$\hat{\beta}_U = \sum_{k=0}^{K} \hat{p}_k (\overline{Y}^*_{Rk} - \overline{Y}^*_{Fk}) \quad (8)$$

dengan:

$\hat{p}_k$ = proporsi peserta tes pada kelompok acuan dan kelompok fokus pada subkelompok X = k;

$\overline{Y}^*_{Rk}$ = rerata tersesuaikan (*adjusted mean*) untuk kelompok acuan pada skor subtes pemadanan X = k;

$\overline{Y}^*_{Fk}$ = rerata tersesuaikan (*adjusted mean*) untuk kelompok fokus pada skor subtes pemadanan X = k;

Nilai $\overline{Y}^*_{Rk}$ dan $\overline{Y}^*_{Fk}$ pada persamaan di atas dicari dari persamaan berikut (Hsin-Hung Li, Nandakumar & Stout, 1995: 8):

$$\overline{Y}^*_{gk} = \overline{Y}_{gk} + \hat{M}_{gk}(\hat{V}(k) - \hat{V}_g(k)) \quad (9)$$

dengan:

$\overline{Y}_{gk}$ = rerata terobservasi (*observed mean*) pada subtes yang diselidiki dari peserta tes pada kelompok g (kelompok R atau F)

$$\hat{M}_{gk} = \frac{\overline{Y}_{g,k+1} - \overline{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g(k-1)} \quad (10)$$

$\overline{Y}_{g,k+1}$ = rerata terobservasi pada subtes yang diselidiki pada subkelompok satu tingkat di atas subkelompok ke-k pada kelompok g (kelompok R atau F)

$\overline{Y}_{g,k-1}$ = rerata terobservasi pada subtes yang diselidiki pada subkelompok satu tingkat di bawah subkelompok ke-k pada kelompok g (kelompok R atau F)

$$\hat{V}_g(k) = \overline{X}_g + (1 - \frac{\hat{\sigma}^2_{(e|g)}}{\hat{\sigma}^2_{(x|g)}})(k - \overline{X}_g) \quad (11)$$

$\overline{X}_g$ = rerata terobservasi subtes pemadanan pada kelompok g (kelompok R atau F)

$$\hat{\sigma}^2_{(x|g)} = \frac{1}{(J_g - 1)} \sum_{j=1}^{J_g} (X_{gj} - \overline{X}_g)^2 \quad (12)$$

J_g = banyaknya peserta tes pada kelompok g (kelompok R atau F)

X_{gj} = skor pada subtes pemadanan pada peserta tes ke-j pada kelompok g (kelompok R atau F)

$$\hat{\sigma}^2_{(e|g)} = \sum_{i=1}^{n} U_{ig}(1 - U_{ig}) \quad (13)$$

n = banyaknya butir soal pada subtes pemadanan

U_{ig} = proporsi jawaban benar pada kelompok g (kelompok R atau F) untuk subtes pemadanan pada butir soal ke-i

$$\hat{V}(k) = \frac{1}{2}(\hat{V}_R(k) + \hat{V}_F(k)) \quad (14)$$

Jika $\hat{\beta}_U > 0$ dan signifikan, maka butir atau butir-butir soal yang bersangkutan terdeteksi *DIF* yang menguntungkan kelompok acuan. Sebaliknya, jika $\hat{\beta}_U < 0$ dan signifikan, maka butir atau butir-butir soal yang bersangkutan terdeteksi *DIF* yang menguntungkan kelompok fokus.

Uji statistik untuk menguji hipotesis nol (ketiadaan *DIF* atau *DTF*) menggunakan uji B. Formula B menurut Hsin-Hung Li, Nandakumar & Stout (1995: 6) dinyatakan dengan:

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)} \quad (15)$$

dengan $\hat{\sigma}(\hat{\beta}_U)$ adalah estimasi kesalahan baku $\hat{\beta}_U$ yang diberikan oleh formula berikut (Hsin-Hung Li, Nandakumar & Stout, 1995: 6):

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=0}^{K} \hat{p}_k^2 \left(\frac{1}{J_{Rk}} \hat{\sigma}^2(Y \mid k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y \mid k, F) \right) \right)^{\frac{1}{2}} \quad (16)$$

dengan:

J_{Rk} = banyaknya peserta tes pada kelompok acuan pada subkelompok k

J_{Fk} = banyaknya peserta tes pada kelompok fokus pada subkelompok k

$\hat{\sigma}^2(Y|k,g)$ = variansi pada skor subtes yang diselidiki pada subkelompok k pada kelompok g (acuan atau fokus)

Statistik uji B pada persamaan (15) berdistribusi normal baku.

Penggunaan metode regresi logistik diperkenalkan pertama kali oleh Swaminathan dan Rogers (1990). Metode ini merupakan metode berdasarkan teori tes klasik yang juga populer di samping metode Mantel-Haenszel dan metode SIBTEST (Embretson & Reise, 2000: 251). Jika metode Mantel-Haenszel dan metode SIBTEST didesain untuk hanya mendeteksi *DIF* uniform, metode regresi logistik dapat dipakai untuk mendeteksi *DIF* uniform dan *DIF* tidak uniform sekaligus.

Pada metode regresi logistik, peluang seseorang menjawab benar suatu butir soal mempunyai bentuk logistik berikut (Swaminathan & Rogers, 1990: 363):

$$P(u=1)=\frac{e^z}{1+e^z} \quad (17)$$

dengan $P(u=1)$ menyatakan peluang peserta menjawab benar suatu butir soal tertentu. Pada metode ini, karena dicari perbedaan antar kelompok (yang menyatakan adanya *DIF* uniform) dan interaksi antara keanggotaan kelompok dan kemampuan peserta tes (yang menyatakan *DIF* tidak uniform), maka z dinyatakan dalam bentuk berikut (Swaminathan & Rogers, 1990: 363):

$$z=\delta+\tau_1 G+\tau_2\theta+\tau_3(G\theta) \quad (18)$$

dengan:

θ = kemampuan peserta tes;

G = kelompok peserta tes, yang dapat diberi kode 1 (untuk kelompok fokus) atau 2 (untuk kelompok acuan).

Jika kemampuan peserta dinyatakan dengan skor total yang diperoleh peserta tes, maka z dinyatakan dalam bentuk berikut (Camilli & Shepard, 1994: 125):

$$z = \delta + \tau_1 G + \tau_2 X + \tau_3 (GX) \quad (19)$$

dengan:

X = skor total yang diperoleh peserta tes;

G = kelompok peserta tes, yang dapat diberi kode 1 (untuk kelompok fokus) atau 2 (untuk kelompok acuan).

Melihat persamaan (19), jika τ_1 signifikan, maka peluang mendapat jawaban benar berbeda untuk kelompok acuan dan kelompok fokus. Jika τ_3 signifikan, maka hal ini menunjukkan adanya *DIF* tidak uniform. Jika τ_3 tidak signifikan tetapi τ_1 signifikan, maka hal ini menunjukkan adanya *DIF* uniform. Jika τ_3 dan τ_1 keduanya tidak signifikan, maka hal ini menunjukkan butir yang bersangkutan tidak memuat *DIF*. Jika τ_1 signifikan dan $\tau_1 > 0$, maka butir yang diselidiki terdeteksi *DIF* yang menguntungkan kelompok acuan, jika $\tau_1 = 0$, butir yang diselidiki tidak memuaf *DIF*, dan jika τ_1 signifikan dan $\tau_1 < 0$, maka butir yang diselidiki terdeteksi *DIF* yang menguntungkan kelompok fokus.

Pengujian signifikansi τ_k dilakukan dengan bantuan paket program SPSS 10.0 *for Windows* sebagai berikut (Field, 2000: 181):

$$W = \frac{\tau_k^2}{[SE(\tau_k)]^2} \quad (20)$$

Statistik uji W pada persamaan (20) berdistribusi khi-kuadrat dengan derajat kebebasan 1.

Pada dasarnya metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik yang disampaikan di muka adalah metode pendeteksian *DIF* yang berdasarkan uji statistik (*statistically DIF detection method*). Prinsip dasar metode-metode tersebut mengikuti prinsip dasar uji statistik pada umumnya, yaitu sebagai berikut. Jika H_0 ditolak, maka butir

soal yang bersangkutan terkena *DIF*. Sebaliknya, jika H_0 diterima (tidak ditolak), maka butir soal yang bersangkutan tidak terkena *DIF*.

Terkait dengan penolakan atau penerimaan H_0 dalam deteksi *DIF* dapat muncul kesalahan, yang disebut kesalahan tipe I dan kesalahan tipe II. Kesalahan tipe I terjadi pada sebuah butir soal oleh suatu metode, jika butir soal tersebut terdeteksi sebagai butir *DIF* yang menguntungkan kelompok tertentu oleh metode tersebut, pada hal butir soal tersebut sebenarnya bukan butir *DIF* yang menguntungkan kelompok tersebut. Kesalahan tipe II terjadi pada sebuah butir soal oleh suatu metode, jika butir soal tersebut tidak terdeteksi sebagai butir *DIF* pada suatu metode, pada hal seharusnya butir tersebut merupakan butir *DIF*. Di sisi lain, sebuah butir soal terdeteksi *DIF* secara benar oleh suatu metode, jika butir soal tersebut terdeteksi sebagai butir *DIF* yang menguntungkan kelompok tertentu oleh metode tersebut dan butir soal tersebut merupakan butir *DIF* yang menguntungkan kelompok tersebut itu.

Dari hal-hal terakhir muncul keingintahuan untuk membandingkan ketepatan metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik. Kecuali pembandingan itu, juga dipertanyakan pada kandungan *DIF* berapa persen suatu metode tidak melakukan kesalahan. Dengan demikian, masalah penelitian dirumuskan sebagai berikut: (a) bagaimana urutan ketepatan pendeteksian *DIF* antara metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik, (b) pada kandungan *DIF* berapa persen, metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik tersebut tidak melakukan kesalahan.

Metode Penelitian

Studi simulasi didesain sebagai berikut: (a) banyaknya peserta tes adalah 1000 peserta tes untuk kelompok acuan dan 1000 peserta tes untuk kelompok fokus dan (b) tes yang dipelajari adalah tes pilihan ganda dengan 5 pilihan jawaban.

Perbandingan ketepatan antar-metode juga dipelajari dari tiga kelompok tes, yaitu kelompok pertama yang terdiri dari 20 buah tes yang

masing-masing terdiri dari 25 butir soal, kelompok kedua yang terdiri dari 40 buah tes yang masing-masing terdiri dari 50 butir soal, dan kelompok ketiga yang terdiri dari 60 buah tes yang masing-masing terdiri dari 75 butir soal. Desain muatan DIF pada setiap kelompok adalah sebagai berikut. Muatan *DIF* pada tes ke-i adalah 4i% pada kelompok pertama, 2i% pada kelompok kedua, dan $\frac{4}{3}$i% pada kelompok ketiga.

Oleh karena penelitian dilakukan melalui simulasi, maka diperlukan pembuatan data (*data generation*). Data yang diperlukan adalah data mengenai respons peserta tes, yang berwujud nilai 1 (benar) atau nilai 0 (salah) pada setiap butir soal, untuk 1000 peserta tes kelompok acuan dan 1000 peserta tes kelompok fokus.

Pembuatan data dilakukan melalui sebuah program dengan bahasa pemrograman Fortran menggunakan Microsoft Fortran Power Station 4.0. Pada dasarnya, algoritma pembuatan data tersebut adalah sebagai berikut: (1) untuk setiap peserta tes, pada kelompok acuan maupun pada kelompok fokus, ditentukan kemampuan laten (*latent ability*) θ; (2) untuk setiap butir soal, pada kelompok acuan dan fokus, ditentukan nilai parameter a (daya beda), b (tingkat kesulitan), dan c (*pseudo guessing*) menurut teori respons butir; (3) untuk setiap peserta pada setiap butir soal dihitung besarnya P(θ) sesuai dengan model logistik tiga parameter, yaitu $P(\theta) = c + \frac{1-c}{1+e^{-1,7a(\theta-b)}}$; (d) berdasarkan nilai $P(\theta)$ yang diperoleh pada langkah (1), ditentukan respons peserta tes, yang berwujud nilai 0 atau 1, dengan cara seperti yang dilakukan oleh Hambleton dan Rovinelli (1973).

Pada pembuatan data di atas, nilai parameter a, b, dan c diambil sedemikian rupa sehingga terdapat muatan *DIF* dengan persentase seperti yang diinginkan. Nilai paramater a, b, dan c pada data ini untuk selanjutnya diasumsikan sebagai nilai parameter yang sebenarnya ada pada populasi dan dipakai untuk menentukan adanya *true DIF* (*DIF* yang sebenarnya). Penghitungan kesalahan tipe I dan kesalahan tipe II serta banyaknya butir soal yang terdeteksi secara benar oleh suatu metode dilakukan dengan membandingkan butir yang terdeteksi *DIF* oleh suatu metode dengan *true DIF* tersebut.

Pada pembuatan data, untuk setiap butir, diambil nilai-nilai parameter sebagai berikut. Untuk butir yang terkena *DIF*, nilai-nilai parameter untuk kelompok acuan (*reference group*) adalah: a = 1,60, b = −0,50, dan c = 0,10 dan nilai-nilai parameter untuk kelompok fokus (*focal group*) adalah a = 1,60, b = 0,50, dan c = 0,10. Untuk butir yang tidak terkena *DIF*, nilai-nilai parameter untuk kelompok acuan adalah: a = 1,60, b = 0,00, dan c = 0,10 dan nilai-nilai parameter untuk kelompok fokus adalah a = 1,60, b = 0,00, dan c = 0,10. Dengan demikian, pada simulasi ini, semua true *DIF* dirancang sebagai *DIF* yang menguntungkan kelompok acuan.

Setelah data dibuat, maka berturut-turut terhadap semua butir soal dilakukan pendeteksian *DIF* dengan metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik. Untuk menentukan butir-butir *DIF* dengan metode Mantel-Haenszel dan metode SIBTEST dibuat program komputer dengan bahasa pemograman Fortran, dan untuk menentukan butir-butir *DIF* dengan metode regresi logistik digunakan SPSS 10.0 *for Windows*. Dengan adanya *true DIF* pada setiap tes, maka pada setiap metode deteksi *DIF* dapat ditentukan banyaknya butir soal yang terdeteksi *DIF* secara benar, banyaknya butir soal yang terkena kesalahan tipe I, dan banyaknya butir soal yang terkena kesalahan tipe II.

Perbedaan ketepatan antarmetode dilihat dari banyaknya butir terkena kesalahan, baik untuk kesalahan tipe I maupun kesalahan tipe II, yang dihasilkan oleh setiap metode. Data dianalisis secara deskriptif dengan melihat banyaknya kesalahan yang dilakukan pada setiap metode. Urutan ketepatan ditentukan oleh urutan banyaknya butir yang terkena kesalahan. Semakin sedikit kesalahan yang dilakukan oleh metode, semakin tepat metode tersebut.

Hasil Penelitian dan Pembahasan

Setelah dilakukan simulasi, banyaknya butir soal yang terkena kesalahan, baik kesalahan tipe I dan kesalahan tipe II, dapat dilihat pada Tabel 2 (untuk panjang 25 butir), Tabel 3 (untuk panjang 50 butir), dan Tabel 4 (untuk panjang 75 butir).

Tabel 2. Banyaknya Butir Soal yang Terkena Kesalahan (Baik Tipe I maupun Tipe II) Berdasarkan Metode Deteksi *DIF* untuk Tes Kelompok Pertama (banyak peserta tes 1000 pada setiap kelompok, panjang tes 25)

Tes ke	KD (%)	Mantel-Haenszel	SIBTEST	Regresi Logistik
Tes ke-1	4	0	0	0
Tes ke-2	8	0	0	0
Tes ke-3	12	17	0	0
Tes ke-4	16	21	0	0
Tes ke-5	20	20	0	0
Tes ke-6	24	19	0	0
Tes ke-7	28	18	0	1
Tes ke-8	32	17	0	6
Tes ke-9	36	16	0	15
Tes ke-10	40	15	0	17
Tes ke-11	44	14	0	20
Tes ke-12	48	13	0	23
Tes ke-13	52	12	0	25
Tes ke-14	56	11	0	25
Tes ke-15	60	10	0	25
Tes ke-16	64	9	0	25
Tes ke-17	68	8	0	25
Tes ke-18	72	7	0	25
Tes ke-19	76	12	0	25
Tes ke-20	80	19	0	25
Jumlah		248	0	282

Keterangan: KD = kandungan *DIF*

Tabel 3. Banyaknya Butir Soal yang Terkena Kesalahan (Baik Tipe I maupun Tipe II) Berdasarkan Metode Deteksi *DIF* untuk Kelompok Kedua (banyak peserta tes 1000 pada setiap kelompok, panjang tes 50)

Tes ke	KD (%)	Mantel-Haenszel	SIBTEST	Regresi Logistik
Tes ke-1	2	0	0	0
Tes ke-2	4	0	0	0
Tes ke-3	6	0	0	0
Tes ke-4	8	0	0	0
Tes ke-5	10	2	0	7
Tes ke-6	12	30	0	38
Tes ke-7	14	41	0	45
Tes ke-8	16	42	0	44
Tes ke-9	18	41	0	43
Tes ke-10	20	40	0	42
Tes ke-11	22	39	0	41
Tes ke-12	24	39	0	40
Tes ke-13	26	37	0	39
Tes ke-14	28	36	0	38
Tes ke-15	30	35	0	37
Tes ke-16	32	34	0	36
Tes ke-17	34	33	0	35
Tes ke-18	36	32	0	34
Tes ke-19	38	31	0	33
Tes ke-20	40	30	0	32
Tes ke-21	42	29	0	31
Tes ke-22	44	29	0	30
Tes ke-23	46	27	0	30
Tes ke-24	48	26	0	28
Tes ke-25	50	25	0	27
Tes ke-26	52	24	0	26
Tes ke-27	54	23	0	25
Tes ke-28	56	22	0	24
Tes ke-29	58	21	0	23
Tes ke-30	60	20	0	24
Tes ke-31	62	19	0	23
Tes ke-32	64	19	0	23
Tes ke-33	66	18	0	22
Tes ke-34	68	19	0	23
Tes ke-35	70	19	0	22
Tes ke-36	72	19	0	21
Tes ke-37	74	23	0	22
Tes ke-38	76	30	0	29
Tes ke-39	78	36	0	42
Tes ke-40	80	40	0	43
Jumlah		1030	0	1122

Keterangan: KD = kandungan *DIF*

Tabel 4. Banyaknya Butir Soal yang Terkena Kesalahan (Baik Tipe I maupun Tipe II) Berdasarkan Metode Deteksi *DIF* untuk Kelompok Ketiga (banyak peserta tes 1000 pada setiap kelompok, panjang tes 75)

Tes ke	KD	Mantel-Haenszel	SIB-TEST	Regresi Logistik		Tes ke	KD	Mantel-Haenszel	SIB-TEST	Regresi Logistik
Tes ke-1	1,3	0	0	0		Tes ke-31	41,3	44	0	46
Tes ke-2	2,7	0	0	0		Tes ke-32	42,7	43	0	51
Tes ke-3	4,0	0	0	0		Tes ke-33	44,0	42	0	55
Tes ke-4	5,2	0	0	0		Tes ke-34	45,3	41	0	59
Tes ke-5	6,5	0	0	0		Tes ke-35	46,7	40	0	66
Tes ke-6	8,0	0	0	1		Tes ke-36	48,0	39	0	70
Tes ke-7	9,3	7	0	1		Tes ke-37	49,3	38	0	72
Tes ke-8	10,7	18	0	3		Tes ke-38	50,7	37	0	73
Tes ke-9	12,0	39	0	3		Tes ke-39	52,0	36	0	75
Tes ke-10	13,3	54	0	3		Tes ke-40	53,3	35	0	74
Tes ke-11	14,7	58	0	5		Tes ke-41	54,7	34	0	76
Tes ke-12	16,0	60	0	6		Tes ke-42	56,0	33	0	78
Tes ke-13	17,3	61	0	8		Tes ke-43	57,3	32	0	78
Tes ke-14	18,7	61	0	9		Tes ke-44	58,7	31	0	78
Tes ke-15	20,0	60	0	10		Tes ke-45	60,0	30	0	78
Tes ke-16	21,3	59	0	10		Tes ke-46	61,3	29	0	78
Tes ke-17	22,7	58	0	10		Tes ke-47	62,7	29	0	77
Tes ke-18	24,0	57	0	14		Tes ke-48	64,0	28	0	77
Tes ke-19	25,3	56	0	13		Tes ke-49	65,3	27	0	77
Tes ke-20	26,7	55	0	15		Tes ke-50	66,7	27	0	78
Tes ke-21	28,0	54	0	16		Tes ke-51	68,0	27	0	79
Tes ke-22	29,3	53	0	18		Tes ke-52	69,3	26	0	80
Tes ke-23	30,7	52	0	19		Tes ke-53	70,7	29	0	80
Tes ke-24	32,0	51	0	22		Tes ke-54	72,0	32	0	81
Tes ke-25	33,3	50	0	24		Tes ke-55	73,3	44	0	81
Tes ke-26	34,7	49	0	26		Tes ke-56	74,7	40	0	81
Tes ke-27	36,0	48	0	28		Tes ke-57	76,0	45	0	81
Tes ke-28	37,3	47	0	30		Tes ke-58	77,3	50	0	80
Tes ke-29	38,7	46	0	36		Tes ke-59	78,7	53	0	80
Tes ke-30	40,0	45	0	41		Tes ke-60	80,0	58	0	81
Jumlah								2287	0	2590

Keterangan: KD = kandungan *DIF*

Berdasarkan Tabel 2, Tabel 3, dan Tabel 4, maka dapat disimpulkan bahwa metode SIBTEST merupakan metode yang mempunyai ketepatan paling baik, disusul oleh metode Mantel-Haenszel, dan metode regresi logistik.

Hasil simulasi dapat dilihat bahwa tidak ada kesalahan sama sekali yang dibuat oleh metode SIBTEST. Metode Mantel-Haenszel tidak melakukan kesalahan apabila kandungan butir *DIF* pada soal sebesar 8% atau kurang, metode regresi logistik tidak melakukan kesalahan apabila kandungan butir *DIF* pada soal sebesar 24% (untuk panjang 25 butir), sebesar 8% (untuk panjang 50 butir), dan sebesar 7% (untuk panjang 75 butir) atau kurang.

Metode SIBTEST merupakan metode yang lebih tepat dibandingkan dengan tiga metode lainnya, disebabkan oleh hal-hal berikut. Pertama, pada metode SIBTEST variabel pemadanannya berubah dari satu butir soal ke butir soal yang lain, sedangkan pada tiga metode yang lainnya variabel pemadanannya tetap dari satu butir soal ke butir soal yang lain. Kedua, pada metode SIBTEST dilakukan penyesuaian rerata skor pada kelompok acuan dan kelompok fokus pada setiap level subtes pemadanan.

Hal lain yang menyebabkan metode Mantel-Haenszel dan metode regresi logistik tidak setepat metode SIBTEST dalam mendeteksi *DIF* yaitu adanya variabel pemadanan, pada metode Mantel-Haenszel dan regresi logistik terkontaminasi oleh butir-butir soal yang terkena *DIF*. Oleh karena itu, beberapa pakar pengukuran menganjurkan untuk melakukan purifikasi (*purification*) ketika memproses pendeteksian *DIF* (Holland & Thayer, 1988: 141; Camilli & Shepard, 1994: 94), karena semakin tidak reliabel skor total sebagai variabel pemadanan, maka semakin besar terjadi kesalahan tipe I (Penfield & Lam, 2000: 7). Holland dan Thayer menyarankan untuk melakukan purifikasi terhadap variabel pemadanan dengan cara tidak mengikutkan butir-butir yang terkena *DIF* ke dalam variabel pemadanan, kemudian mengulangi lagi proses penentuan *DIF* dengan variabel pemadanan yang baru. Dengan cara purifikasi diharapkan diperoleh deteksi *DIF* yang tepat.

Penelitian simulasi mengenai ketepatan pendeteksian *DIF* juga menunjukkan bahwa semakin banyak butir *DIF* terkandung dalam suatu soal, semakin tidak tepat suatu metode mendeteksi *DIF*. Hasil ini mendukung hasil penelitian Fidalgo dan Mellenberg (2000) pada metode Mantel-Haenszel.

Simpulan

Berdasarkan studi simulasi menggunakan tes dengan 5 alternatif jawaban pada panjang 25 butir, 50 butir, dan 75 butir, serta dengan menggunakan 1000 peserta tes pada kelompok acuan dan 1000 peserta tes pada kelompok fokus, diperoleh simpulan sebagai berikut.

1. Urutan ketepatan metode pendeteksian *DIF* dari tiga metode yang dipelajari, mulai yang paling tepat, adalah metode SIBTEST, metode Mantel-Haenszel, dan metode regresi logistik.
2. Pada kandungan *DIF* paling banyak 8%, metode Mantel-Haenszel tidak melakukan kesalahan.
3. Metode SIBTEST tidak melakukan kesalahan sama sekali pada setiap persentase kandungan *DIF*.
4. Pada kandungan *DIF* paling banyak 24% (untuk panjang 25 butir), paling banyak 8% (untuk panjang 50 butir), dan 7% (untuk panjang butir 75%), metode regresi logistik tidak melakukan kesalahan.

Saran

Dari hasil penelitian ini dapat dikemukakan saran sebagai berikut.

1. Kepada para praktisi pengukuran dan pengujian, jika tidak ada kendala pembuatan program atau telah tersedia paket programnya, disarankan untuk menggunakan metode SIBTEST dalam praktik analisis *DIF*, karena metode SIBTEST tidak melakukan kesalahan sama sekali untuk berbagai macam kandungan *DIF* dalam suatu soal.
2. Jika kandungan *DIF* pada suatu soal paling banyak sekitar 8%, keempat metode, yaitu metode Mantel-Haenszel, metode SIBTEST, dan metode regresi logistik.
3. Jika kandungan *DIF* cukup besar dan jika menggunakan metode Mantel-Haenszel atau metode regresi logistik, para praktisi pengukuran dan pengujian disarankan untuk melakukan purifikasi terhadap variabel pemadanannya.

Daftar Pustaka

Camilli, S. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. Dalam P. W. Holland & H. Wainer (Eds), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates Publisher.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Marwah, NJ: Lawrence Erlbaum Associates Publisher.

Fidalgo, A. M. & Mellenberg, G. J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*. 5. 43-53. Diambil pada tanggal 10 Mei 2003, dari http://www.mpr-online.dc.

Field, A. (2000). *Discovering statistics using SPSS for windows: Advanced techniques for the beginner*. London: Sage Publications.

Hambleton, R. K. & Rovinelli, R. (1973). *DATAGEN: A Fortran V program for simulation of dichotomously scored, unidimensional item response data*. UMASS.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Dalam H. Wainer & H. I. Braun (Eds), *Test Validity* (pp 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates Publisher.

Hsin-Hung Li, Nandakumar, R., & Stout, W. (1995). Application of SIBTEST in dealing with issues of DIF in the context of multidimensional data. *Paper*. Presented at the annual meeting of the National Council on Measurement in Education, San Fransisco, 19 April, 1995. Diambil pada tanggal 15 Mei 2003 dari http://www.stat.uiuc.edu/stoutlab/papers.

Hulin, C. L., Drasgow, F. & Parson, C. K. (1993). *Item response theory*: *Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Penfield, R. D. & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5-15.

Shealy, R. T. & Stout, W. F. (1993). An item responses theory model for test bias and differential test functioning. Dalam P. W. Holland & H. Wainer (Eds), *Differential Item Functioning*. (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates Publisher.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement. 27*. 361 – 370.